The Impact of Response Structure on the McGurk Effect

A Senior Honors Thesis

Presented in Partial Fulfillment of the Requirements for graduation with distinction in Speech and Hearing Science in the undergraduate colleges of the Ohio State University

By

Patricia Hyatt

The Ohio State University
June 2008

Project Advisor: Dr. Janet M. Weisenberger, Department of Speech and Hearing Science

Abstract

Understanding speech is a complicated process. Once thought to be a unimodal function, the perception of speech has since been demonstrated to be a multimodal process. The compensatory use of visual information during speech perception is easily evidenced by examining communication in compromised listening situations, like in a noisy restaurant. In such an environment, listeners use visual cues to help understand speech and fill in the missing pieces of auditory information. What is more interesting is that people use visual cues to process speech even when the auditory signal is perfect (McGurk and MacDonald, 1976). The combining of auditory and visual cues during speech perception is termed audio-visual integration.

The subjects in McGurk and MacDonald's study were presented with "discrepant" auditory and visual stimuli and perceptual responses were recorded. The results showed that when a listener was presented, for example, with the audio syllable /ba/ simultaneously with the visual syllable /ga/, the most frequently reported response was /da/, a fusion of the two sounds. This phenomenon has been termed "the McGurk Effect" and has been used to explore audio-visual integration.

Although reports in the literature indicate a substantial degree of McGurk-type integration by both normal-hearing and hearing-impaired subjects (e.g., Grant and Seitz, 1998), previous studies in our laboratory have not found such a high incidence of McGurk integration. One possible explanation for this difference is the fact that previous work in our laboratory employed an open-set response task, in which respondents were not given a fixed set of response options, whereas studies such as that of Grant and Seitz employed a closed-set response task.

The present study explored how the type of response task might influence the McGurk Effect. In the present study, 20 normal-hearing participants were presented with audiovisual syllables featuring a degraded auditory component. Half of the subjects were tested with a closed-set task, and the rest were tested with an open-set task. Results indicated significantly higher incidence of McGurk-type integration for subjects tested with the closed set response task. These findings are discussed in terms of their implications for the development of aural rehabilitation programs for hearing-impaired persons.

Acknowledgments

This study would not have been possible without the guidance of my advisor Dr. Janet Weisenberger, the help of Natalie Felepelle, or the participation of all my subjects. Thank you!

An ASC Undergraduate Scholarship and an SBS Undergraduate Research Scholarship supported this project.

Table of Contents

Chapter 1: Introduction and Literature Review

Audiovisual speech perception was long thought to occur only in situations in which the auditory signal is compromised in some way, as in a noisy environment. In such situations, the listener uses visual cues to help supplement for loss of information from the auditory signal. However, McGurk and MacDonald (1976) conducted a study showing that people use visual cues to process speech even when there is a perfect auditory signal.

McGurk and MacDonald used video tapes with different auditory cues dubbed over the original auditory information to present simultaneous auditory and visual stimuli representing different syllables. The results of their study showed that when a listener was presented with the audio stimulus /ba/ and the visual stimulus /ga/, the reported response was /da/, a fusion of the two sounds. This fusion has been termed "the McGurk Effect." It was also found that in reversing the pair of phonemes, an audio /ga/ with a visual /ba/, subjects reported hearing /bga/. Interpreting the many aspects of audiovisual integration requires knowledge of the information conveyed in audio and visual speech cues separately.

Auditory Cues for Speech Perception

There is much information provided through the auditory component of speech. A listener can extract articulatory information about a speech sound, such as its place, manner and voicing. Place of articulation refers to where the sound was articulated in the mouth. Possible places a sounds can be articulated include: palatals (the tongue and the

hard palate), velars (the tongue and the soft palate), palatal-alveolar (the blade of the tongue and the alveolar ridge), alveolar (the tip of the tongue and the alveolar ridge), interdentals (the tongue between the teeth), labiodentals (the lower lip and upper set of teeth), and bilabials (using both lips). Manner of articulation describes how the articulators interact with each other in each of the place settings. There are stops, liquids, glides, fricatives, and affricates. Voicing refers to the action of the vocal folds in the production of the sound. Vibrating vocal folds create a voiced sound, while non-vibrating vocal folds produce a voiceless sound. These aspects of the auditory cue are processed by the brain and allow a listener to understand a speech signal.

The speech signal travels in a waveform, and the waveform has both temporal and spectral aspects that encode the information the listener can extract. With all of the place, manner and voicing information that is contained in the waveform, there is much more information available in the waveform then a listener may actually need to identify the signal. Because of the large amount of this repeated information, the speech signal has been referred to as highly "redundant."

Shannon et al. (1995) studied the redundancy of speech signals by reducing the spectral information they contained. The signals were stripped of certain elements to determine which cues are necessary for a listener to understand the speech signal. The degradation of the speech signals was done in a fashion similar to the way a cochlear implant degrades an auditory cue. Shannon et al. did this by replacing the fine structure of the waveform with noise.

The fine structure was degraded by taking the temporal envelope of the waveform and using broad noise bands to divide the envelope into one, two, three, and four broad

bands of noise. The filters used had cutoff frequencies of 16, 50, 160, and 500 Hz. Combining these conditions in different ways led to 16 different situations that were presented to eight normal hearing listeners using both nonsense syllables and sentences.

Shannon et al. found that as the number of noise bands used increased, so did a listener's ability to understand the auditory signal. It was found that using three broad noise bands resulted in 90% correct identification. The fact that a small number of broad noise bands can be sufficient for a person to understand a degraded auditory signal supports the idea that the speech signal is redundant.

Visual Cues for Speech Integration

Auditory speech perception is influenced by visual cues. Visual Phonemes, or visemes, have been defined by Fisher (1968, cited by Jackson) as "any individual and contrastive visually perceived unit." A viseme can contain more than one speech sound, or may be a single speech sound. The information contained in visemes is primarily information about the place of articulation of a sound. Visemes help listeners to tell the difference between bilabial sounds, such as /p/ in words like pat, and velar sounds, such as /k/ in words like cat. However, visemes do not contain much information about the voicing of a sound, and therefore, sounds like /b,m,p/ may be misunderstood by a speech reader. All of the sounds have a similar visual placement and are thus visually indistinguishable. Although they are all part of a viseme category, they are clearly differentiated by the auditory cues in the speech signal. However, the fact that the phonemes /b,m,p/ have the same visual placement does not secure their place in a viseme group. There are many factors that contribute to forming a group of sounds as visemes.

Talker variations and differences in the environment in which the auditory cue is produced are important factors to defining viseme groups, and the visual contribution to speech perception (Jackson, 1998).

Auditory-Visual Integration Theories

Knowing that both auditory and visual cues contribute to the understanding of speech sounds, many theories of how the two modalities that contribute to speech perception are integrated have been created. Grant (2002) describes two main theories that have emerged, the Fuzzy Logical Model of Perception (FLMP) and the Pre-Labeling Model of Integration (PRE).

The FLMP was used in research by Massaro (1987) and described as a method that starts with the listener processing the different modalities of the incoming speech information independently. A listener in this model processes the auditory signal separately from the visual signal and then compares it with existing knowledge and memories of the listener. The modality that triggers the strongest memory is the signal that will have a greater influence in the understanding of the speech signal. Massaro talked about a multiplicative integration rule which is used to determine, or predict, the performance of a listener's auditory-visual speech perception. However, the predictions that have been obtained using the FLMP have been outperformed by human observers, making it difficult to say the FLMP actually predicts optimum performance levels.

The Pre-Labeling Model of Integration differs from the FLMP because it does not seek to predict the optimal outcome of a listener's auditory-visual speech perception. The PRE model seeks to "label" the incoming bimodal stimuli. The labels are

determined by first using audio-only and video-only response matrices to find an estimate of the unimodal information. Then an optimum combination rule is used to predict how an unbiased listener with no interference across the modalities might process the particular unimodal information. In Grant's studies the predictions of the PRE model have always equaled or exceeded the observed audio-visual integration scores of the listener. The PRE model provided a better fit to the data in Grant and Seitz's study (1998) which looked at the differences across individuals with hearing impairments and their abilities to integrate audio and visual stimuli. Establishing that the PRE model fits the data from Grant and Seitz's study does not necessarily mean that it is a more correct model to use across all studies. It is important to remember that the ability to integrate information is separate from the ability to extract information from the auditory and visual speech cues, and the validity of the derived estimates of auditory-visual integration cannot be based entirely on model fits (Grant, 2002).

The Importance of the McGurk Effect

Many studies have since looked further into the properties of the McGurk effect. Such studies have shown that the McGurk effect is developmentally strengthened with increasing age (McGurk &McDonald, 1976). "Context effects" have been shown that speech perception is influenced by surrounding speech structure (Repp, 1982 as cited in Green, 1998). There are effects cross-culturally and from different languages on the amount of McGurk responses that appear (Massaro et al., 1993 as stated in Green, 1998). All of these studies have helped to understand the underlying mechanisms of the McGurk effect.

Strong McGurk effects have also been observed with hearing-impaired listeners (Grant & Seitz, 1998). Their study focused on trying to measure the differences among individuals in the amount of benefit they gained from audio and visual cues being combined in the speech signal. Grant and Seitz used audio alone, visual alone, and combined both audio and visual stimuli in their procedures. Subjects were tested with and without McGurk type stimuli. The stimuli that paired an auditory /ba/ paired with a visual /ga/, auditory /pa/ with a visual /ka/, auditory /ma/ with a visual /da/, and an auditory /va/ with a visual /da/ were found to elicit strong McGurk responses. The responses, from both normal hearing and hearing impaired individuals showed evidence of McGurk type responses exist, attesting to integration of the auditory and visual stimului.

Recent studies in our laboratory at The Ohio State University have also used the McGurk paradigm to study audio-visual integration more generally (Huffman, 2007; Andrews, 2007). A series of studies explored the different audio and visual components that aid in the integration process, how the integration process can be affected when either audio or visual signals are degraded, the particular characteristics that promote audiovisual integration from certain talkers, and also the individual differences in integration efficiency produced by talkers. Specifically, we are interested in how listeners integrate visual input with degraded auditory signals from which specific types of information have been removed. Surprisingly, these recent studies found very low levels of McGurk-type integration responses. Although it could be hypothesized that the low levels of McGurk integration are attributable to the degree of degradation of the auditory signal, higher levels of McGurk-type integration responses might be expected, given the

results reported for hearing impaired listeners by researchers such as Grant and Seitz (1998) for hearing-impaired listeners. It is possible that differences in the results can be attributed to differences in the acoustic stimulus. The studies by Huffman and Andrews employed stimuli which the spectral fine structure was replaced by noise, but the temporal envelope characteristics were retained. This approach was similar to that used by Shannon et al. (1995) to simulate the stimulus presented by a cochlear implant.

However, a second possible difference between the studies of Huffman and Andrews and that of Grant and Seitz (1998) that may be important is the structure of response options given to subjects. Grant and Seitz used closed set response tasks, in which subjects were asked to choose from specific response options. The previous studies in our laboratory employed an open response structure in which subjects could respond with any syllable. The present study explores the impact of the structure of response task. One group of subjects was given a list of syllables to chose from, a closed set, and the other group of subjects was given an open set and allowed to answer with any syllable. Aas expected, the results of the present study show that higher levels of McGurk-type integration wree elicited by the group of individuals that used the closed set response task.

Chapter 2: Method

Participants

There were 20 people who participated as observers. All participants were undergraduate/graduate university students with reported normal or corrected vision and normal hearing. Ages ranged from 19-26. Nine of the observers were male, eleven were female. Two of the participants were undergraduate students who were majors in Speech and Hearing Science. One student received academic credit for participating in the study, while the other 19 received payment of forty dollars for their time.

There were 5 people who participated as talkers. There were 3 females, 2 males with ages ranging from 20-23. All were native speakers of English.

Interface for Stimulus Presentation

A 20 inch video monitor, placed four feet away from the observer's chair, was used to present visual stimuli. Auditory stimuli were presented using TDH 39 headphones.

Stimuli Selection

A limited set of CVC syllables were presented as stimuli for this study. The set was chose based on the ability of the stimuli to satisfy the following conditions:

1. Pairs of the Stimuli were minimal pairs, that is, they differed only by the initial consonant.

2. All stimuli were accompanied by the vowel /ae/, because it does not exhibit lip rounding or lip extension.

3. Multiple stimuli were used in each category of articulation, including; place (bilabial, alveolar), manner (stop, fricative, nasal), and voicing (voiced, unvoiced).

4. All stimuli were presented without a carrier phrase.

5. Stimuli were known to elicit McGurk-like responses when appropriately chosen pairs of syllables were presented.

Stimuli

For each condition (degraded auditory-only, visual-only, and degraded auditory & visual), the same set of eight stimuli were administered:

Bilabial:

1. mat
2. bat
3. pat

Alveolar:

4. sat
5. zat
6. tat

Velar:

7. gat
8. cat

In the degraded auditory and visual condition, the four following dual-syllable (dubbed) stimuli were used. The first column indicates the visual stimulus, and the second column indicates the auditory stimulus.

1. bat-gat
2. gat-bat
3. cat-pat
4. pat-cat

Stimulus Degrading and Recording

The stimulus set that was used for the present study was recorded using the software program Video Explosion Deluxe. Each of the five talkers was recorded producing each of the eight syllables five times. Their voices were recorded directly onto a computer, through the use of a microphone, which allowed the stimuli to be stored in a .wav format and input into a subroutine in MATLAB 5.3 (Smith, 2002). The MATLAB 5.3 program exchanged the amplitude envelope waveform and the spectral fine structure of the two stimuli, then the signal is filtered into four broad spectral bands. The bandwidths of the four spectral bands were 504 Hz, 1,794 Hz, 5,716 Hz, and 17,604 Hz; these were chosen specifically so that the four channels would provide equal spacing along the distance of the basilar membrane.

Digital videos were then created using the degraded auditory clips and recorded video of the talkers in the program Video Explosion Deluxe. The existing video clips had the auditory clips dubbed onto them; this made it possible to create stimuli that could produce a McGurk-type response. The McGurk type response can be elicited when the

visual stimulus and auditory stimulus are incongruent, for example a visual stimulus of /bat/ would be dubbed over with an auditory stimulus of /gaet/. The ability to place any visual and auditory stimuli together also allowed the auditory signal to be degraded in the stimulus set. The present study paired visual and auditory stimuli produced by the same talker. Three randomized lists of 60 stimulus clips were created for each of the five talkers and compiled into three DVDs using the programs Video Explosion Deluxe and Sonic MY DVD. The present study used a total of 15 DVDs.

Procedure

Testing for this study was conducted in the basement lab of Pressey Hall on the campus of The Ohio State University. Each participant was tested individually inside a sound-attenuating booth, with the door shut. Participants sat facing the window of the booth so that they could view a 50cm television monitor that was placed outside the booth on the other side of the double glass window. Subjects sat at a distance of approximately 4 feet from the monitor. Auditory stimuli were presented to the subjects through TDH 39 headphones. Subjects' responses were transmitted to the examiner outside of the booth using an intercom system.

Participants in the Closed Set group were given a list of 17 syllables and directed that they would be presented with different talkers under auditory-alone, visual-alone, and auditory+visual conditions where the talkers would be saying one of the words on the list of terms. They were also told that all the syllables and syllable combinations they would be hearing would end with "at" and differ only in the initial consonants. Participants belonging to the Open Set group were given the same instructions, but they

were not given a list of words to choose from. Subjects in the Open Set group were told that they may hear or see any combination of sounds, even sounds that did not exist in the English language, but that the syllable presented to them would always end in “at.”

Each observer was presented with 3 videos each of 5 different talkers, making a total of 15 DVDs. Each DVD contained 60 stimulus presentations that were randomly ordered. The observer watched one DVD of each talker in an auditory-only, visual-only, and auditory+visual condition. Auditory-only stimuli were presented via headphones, and the monitor of the television was turned off. For visual-only presentation the headphones were turned off and the television monitor was on. Both the headphones and television monitor were on for auditory+visual stimulus presentation. In an auditory+visual presentation the stimuli set consisted of 60 stimuli, but 30 of the stimuli were “same trials” and 30 were “different trials.” The “same trials” allowed for a percent correct to be computed, and the “different trials” presented auditory and visual stimuli that conflicted with each other and presented opportunities for the observer to create a McGurk-type response. The examiner listened over an intercom and wrote down the subjects responses to the stimuli; rest periods were encouraged to minimize both observer and examiner fatigue.

Chapter 3: Results and Discussion

Percent Correct Performance

Figure 1 measures the percent correct identification performance under auditory-only, visual-only, and auditory+visual modalities for both open and closed response set conditions. Results were averaged across subjects in each group. Contrary to our expectations, the results across the two groups in the visual-only modality were nearly the same; 29% correct in a closed condition response and 30% correct in an open condition response set. Similarly, auditory only conditions led to a percent correct performance of 45% in a closet set response task and 44% in the open set condition. Audio-visual integration is shown in both closed and open condition of response task. The integration is shown by the increase in percent correct from auditory-only to auditory+visual presentation. A small difference between response tasks was observed here. The closed set group had an auditory-only percent correct of 45% and increased by 22% to have a 67% correct response rate in the auditory+visual condition. The open set group had an auditory-only percent correct of 44% and increased by 16% to have a 60% correct response rate in the auditory+visual condition.

A two factor mixed model analysis of variance was performed to determine the significance of any differences in the data. There was no significant effect of response set, $F(1,18)=.671$, $p=.423$. Not surprisingly, there was a significant main effect of presentation condition, $F(2,36)=202.916$, $p < .001$. There was an insignificant interaction effect, $F(1,18)=4.25$, $p= .054$.

Talker Effects

Figures 2 and 3 show performance under each testing condition for individual talkers. Auditory+visual testing conditions are slightly better for every talker in the closed set response condition than the open set response condition. However, auditory-only and visual-only performances appear relatively unaffected by response condition. As found in previous studies with their talkers, there are substantial differences among talkers. LG has the best overall levels of performance, but produces less integration than talker KS.

McGurk Data

Figure 4 shows the percent that subjects chose a particular modality of presentation of the stimuli. To elicit a McGurk-type response, it is necessary that the auditory and visual signals in a stimulus are not signals for the same syllable; they must be discrepant. Because of the discrepant signals in the same stimulus there is no correct answer. Instead, subjects may choose the signal delivered by one modality over the other, or they might integrate the modalities. Figure 4 plots the percentage of times that a subject chose each modality of the stimulus, be it the auditory component, visual component, or if a response that did not match either modality entirely. The figure shows the reliance on specific sensory modalities in response condition. There is a 12% reliance on the visual modality in a closed set condition and a14% reliance in the open set condition. There is also little change with auditory only responses, 33% in the closed set condition and 36% in the open set condition. There are quite a few “other” responses in both conditions (55% in closed set conditions and 51% in open set conditions). Tthese

responses are importantbecause some percentage of theese may reflect McGurk-type responses. The “other” category is analyzed further in Figure 5.

To determine any significant differences in the data in figure 4 a two factor mixed model analysis of variance was performed. There was a significant difference in the response modality, $F=(2,36)=60.5411$, $p < .001$. There was no significant difference found between closed set and open set response conditions, $F(1, 18)=.463$, $p=.505$. There was also no significant interaction effect, $F(1,18)=.885$, $p=.395$.

Figure 5 shows the percent of the “other” category of responses that can be counted as McGurk-type integration responses. Fusion occurs when the place of articulation is an intermediate location between the two places of articulation for the two modalities presented in the discrepant stimuli. For example, a visual syllable “pat” paired with an auditory syllable “cat” listeners would report hearing “tat.” Combination responses include both of the places of articulation for the visual and auditory signals in the discrepant stimuli. Using the same syllables, visual “pat” and auditory “cat” a listener would report hearing “pcat.” Thus, fusion and combination responses are both examples of McGurk-type integration. In Figure 5 it is apparent subjects more often used combination responses in the closed set response condition than in the open set response condition.

To analyze the difference between the responses and the response condition, a two factor mixed model analysis of variance was performed on the data for Figure 5. The main effects (fusion, combination, neither) were all shown to be significantly different from one another, $F(2,36)=34.681$, $p<.001$. Between closed set and open set response groups there was found to be no significant difference, $F(1,18)=1.169$, $p=.294$.

Interaction effects were also found insignificant, $F(1,18)=.017$, $p=626$. Further statistical analysis focused on comparing specific response type. This was done using a between group t-Test. The effect of response condition specifically for combination McGurk-type responses was found significant, $t(18)=3.459$, $p<.001$. For fusion responses $t(18)=-1.073$, $p=.994$, the effects of response set are not significant.

The total number of McGurk type responses is shown in Figure 6. A between group t-test was performed to determine if response set had a significant effect on the overall amount of integration that occurred fo rthe discrepant stimuli. Results showed that even though there wa a higher total percentage of McGurk (integrated) responses in the closed set response task, the difference was not significant . $t(18)=1.597$, $p=.375$.

Chapter 4: Summary and Conclusion

Grant and Seitz (1998) as well as many others, have found larger amounts of McGurk-type responses than our laboratory had found in similar studies. The results of the present study, particularly in Figure 5, show that when a closed set response task is employed, subjects are less likely to answer with a non-McGurk response. It is also interesting to note that closed-set response tasks elicit substantially more combination responses. Previous studies in this lab have employed open set response tasks, while Grant and Seitz used closed set response tasks. This may not be the only reason for the differences in the percentage of integration, but it cannot be ignored as a factor.

Results of this study imply that it may be beneficial for aural rehabilitation programs to incorporate closed set response tasks into the curriculum. Results show that while an overall percent correct may not go up significantly between open and closed set response groups, the level of integration does increase. Successful communication uses integration of both auditory and visual speech cues and should be a part of the goal of aural rehabilitation programs. This is why improving integration skills should be included in the aural rehabilitation program, and progress should not just be measured in percent correct responses.

Chapter 5: References

Andrews, B. (2007). Auditory and visual information facilitation speech integration (Senior Honors Thesis, The Ohio State University, 2007).

Green, K.P. (1998). The use of auditory and visual information during phonetic processing: Implications for theories of speech perception. Campbell, R., Dodd, B., and Burnham, D. (eds.), *Hearing eye II: Advances in the psychology of speechreading and auditory-visual speech*. East Sussex, UK: Psychology Press, Ltd.

Grant, K.W. & Seitz, P.F. (1998). Measures of auditory-visual integration in nonsense syllables and sentences. *The Journal of the Acoustical Society of America*, 04, (4), 2438-2449.

Grant, K.W. (2002). Measures of auditory-visual integration for speech understanding: A theoretical perspective (L). *The Journal of the Accoustical Society of America*, 112 (1), 30-33.

Huffman, C.(2007) The role of auditory information in audiovisual speech integration (Senior Honors Thesis, The Ohio State University, 2007).

Jackson, P. L. (1988). The theoretical minimal unit for visual speech perception: Visemes and coarticulation. The Volta Review, 90 (5), 99-114.

McGurk, H. & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264, 746748.\.

Shannon, R.V., Aeng, F.G., Wygonski, J., Ekelid, M. (1995). Speech Recognition with primarily temporal cues. Science, 270, 303-304.

Smith, Z.M., Oxenham, A.O., and Delgutte, B. Chimaeric sounds reveal dichotomies in auditory perception. *Nature*, 2002; 416:87-90.

List of Figures

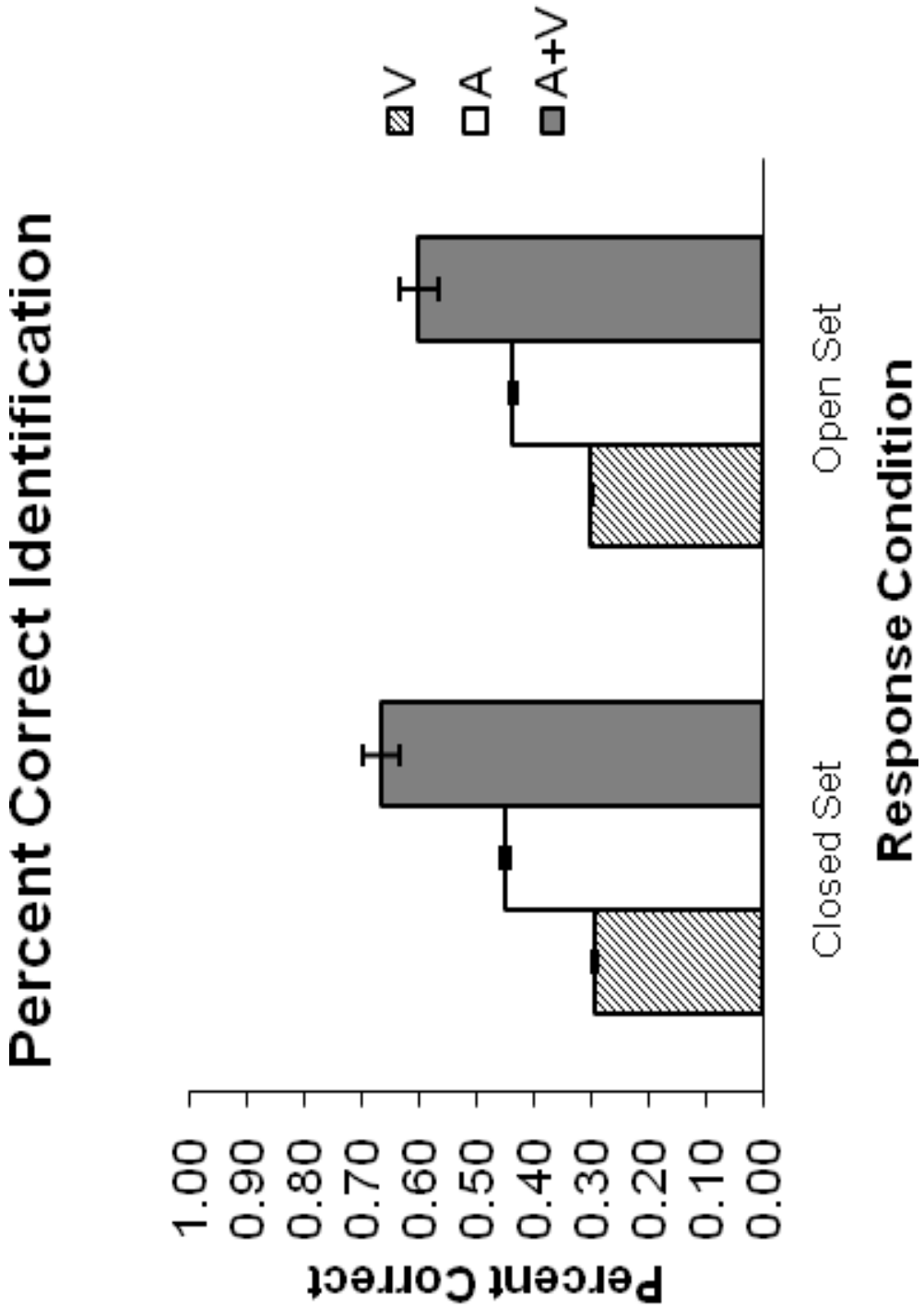
Percent Correct Identification
V
A
A+V
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
Percent Correct
Closed Set
Open Set
Response Condition

Figure 1

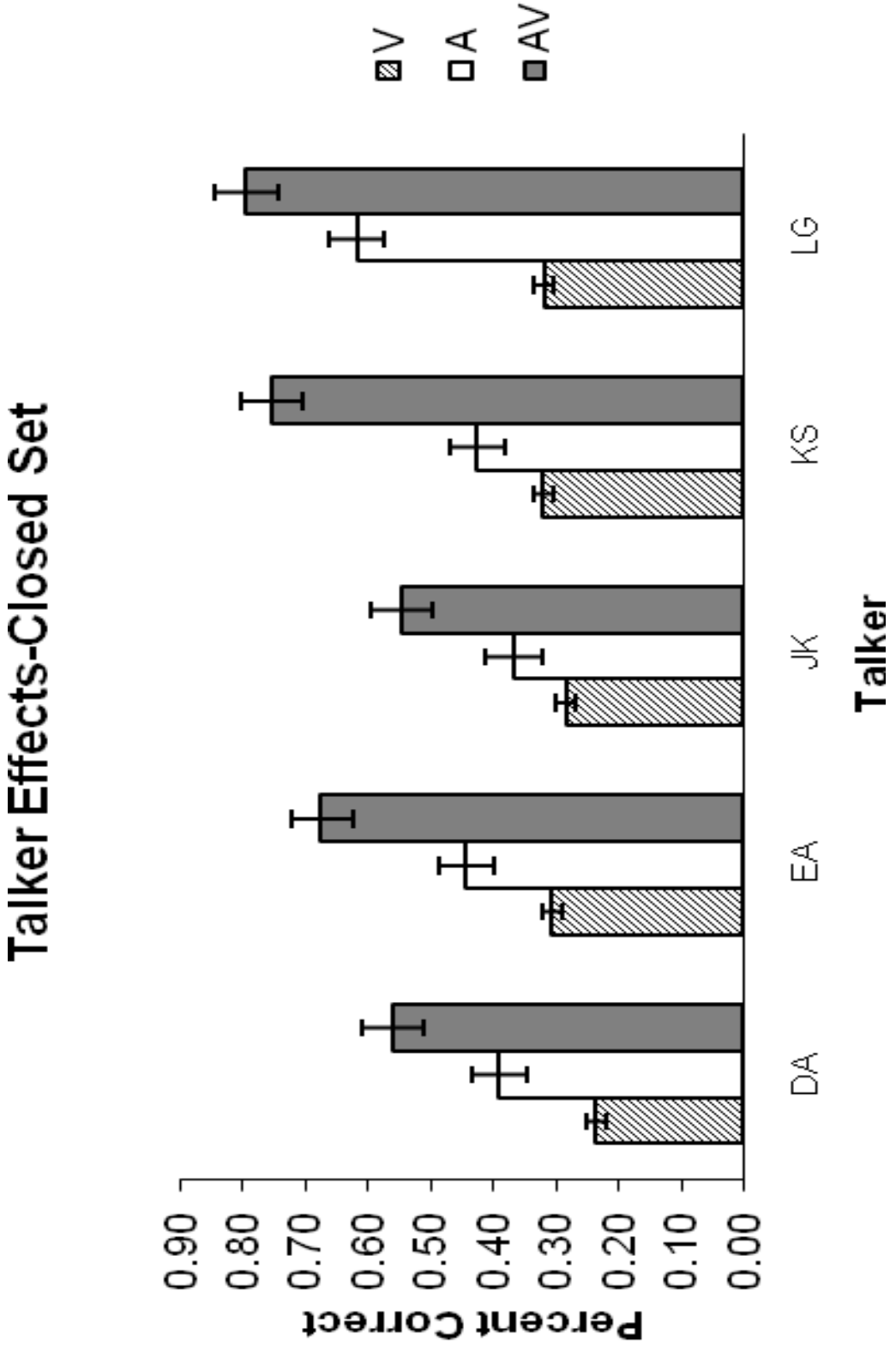
Talker Effects-Closed Set
V
A
AV
Percent Correct
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
DA
EA
JK
KS
LG
Talker

Figure 2

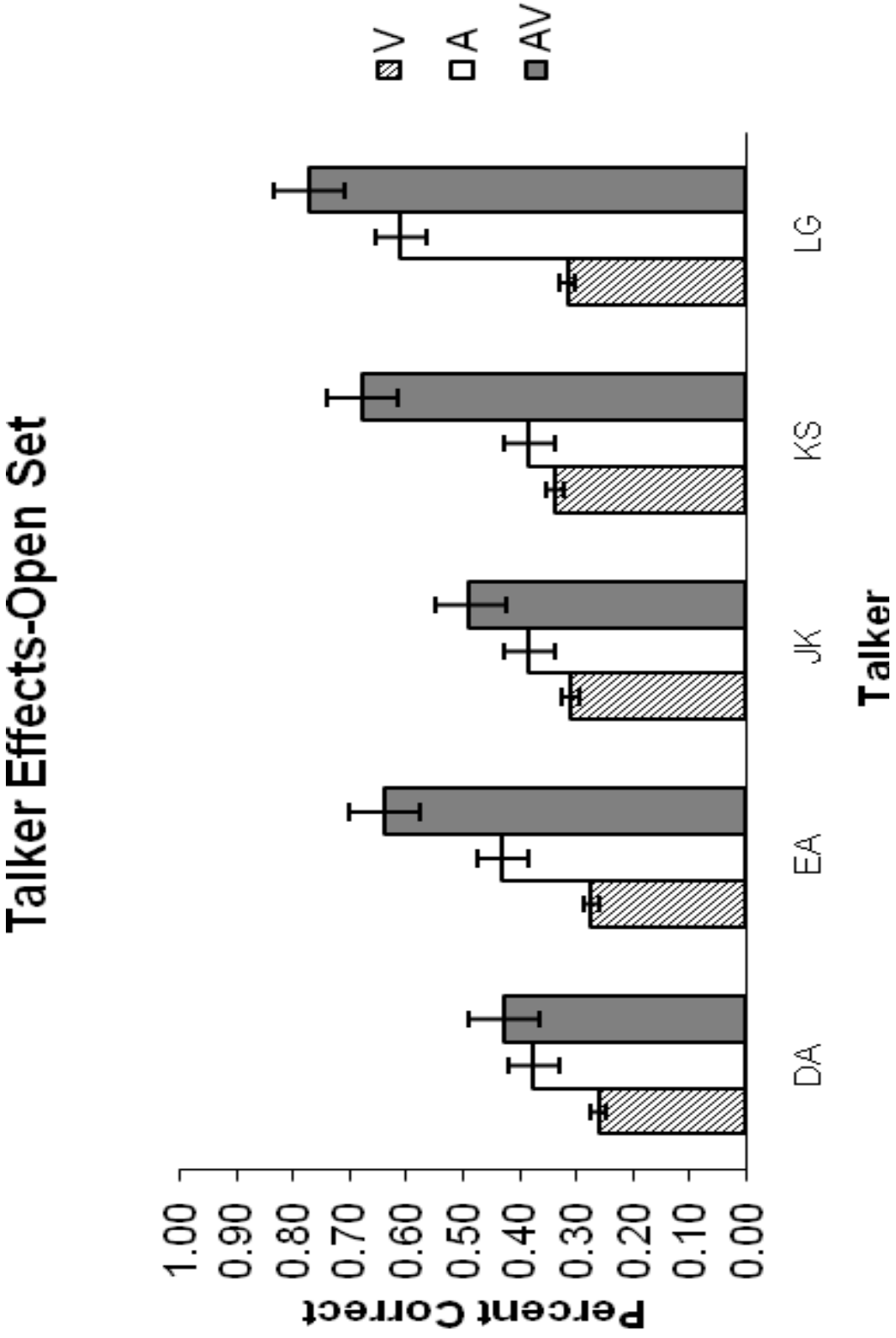
Talker Effects-Open Set
V
A
AV
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
Percent Correct
DA
EA
JK
KS
LG
Talker

Figure 3

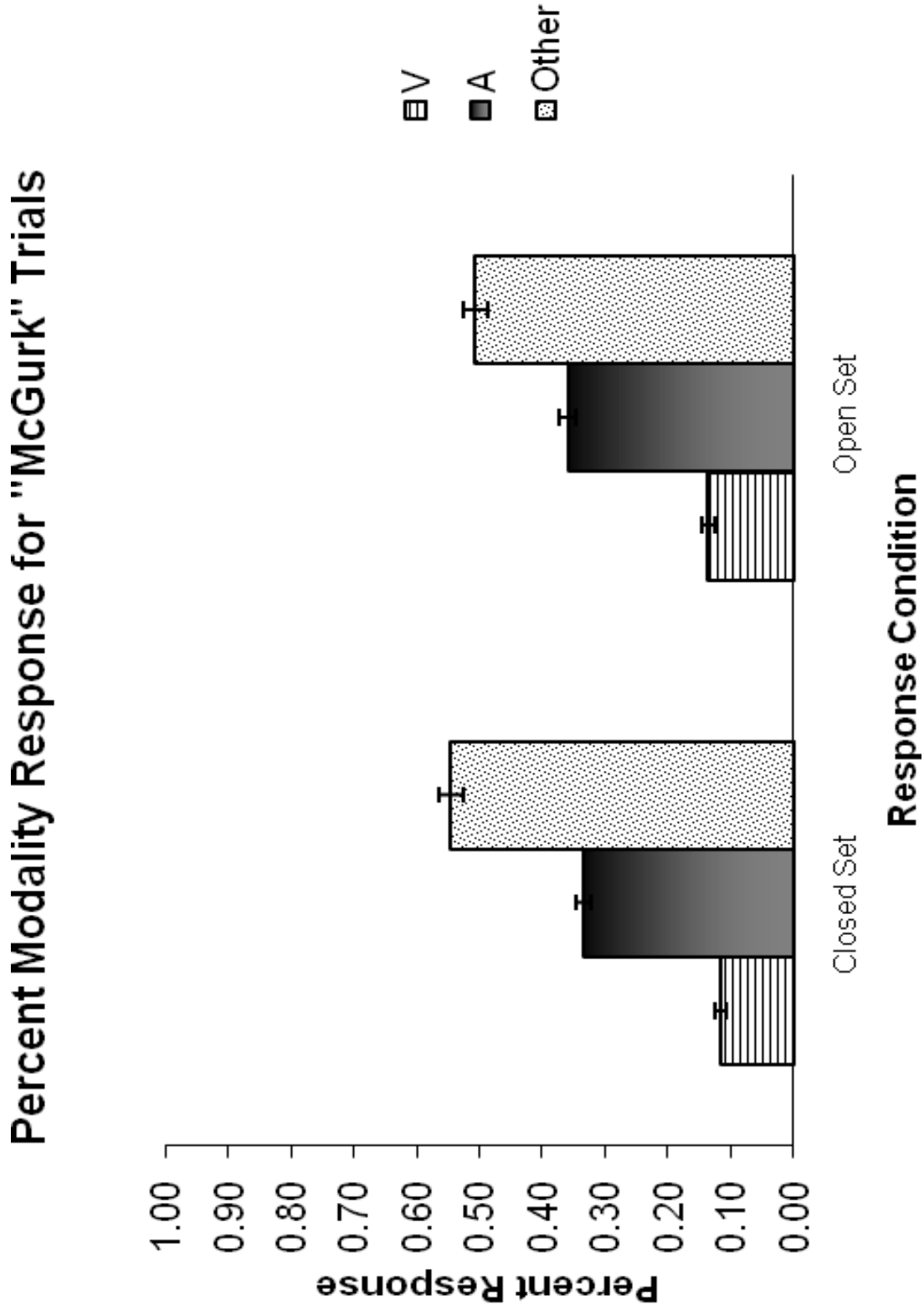
Percent Modality Response for "McGurk" Trials
V
A
Other
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
Percent Response
Closed Set
Open Set
Response Condition

Figure 4

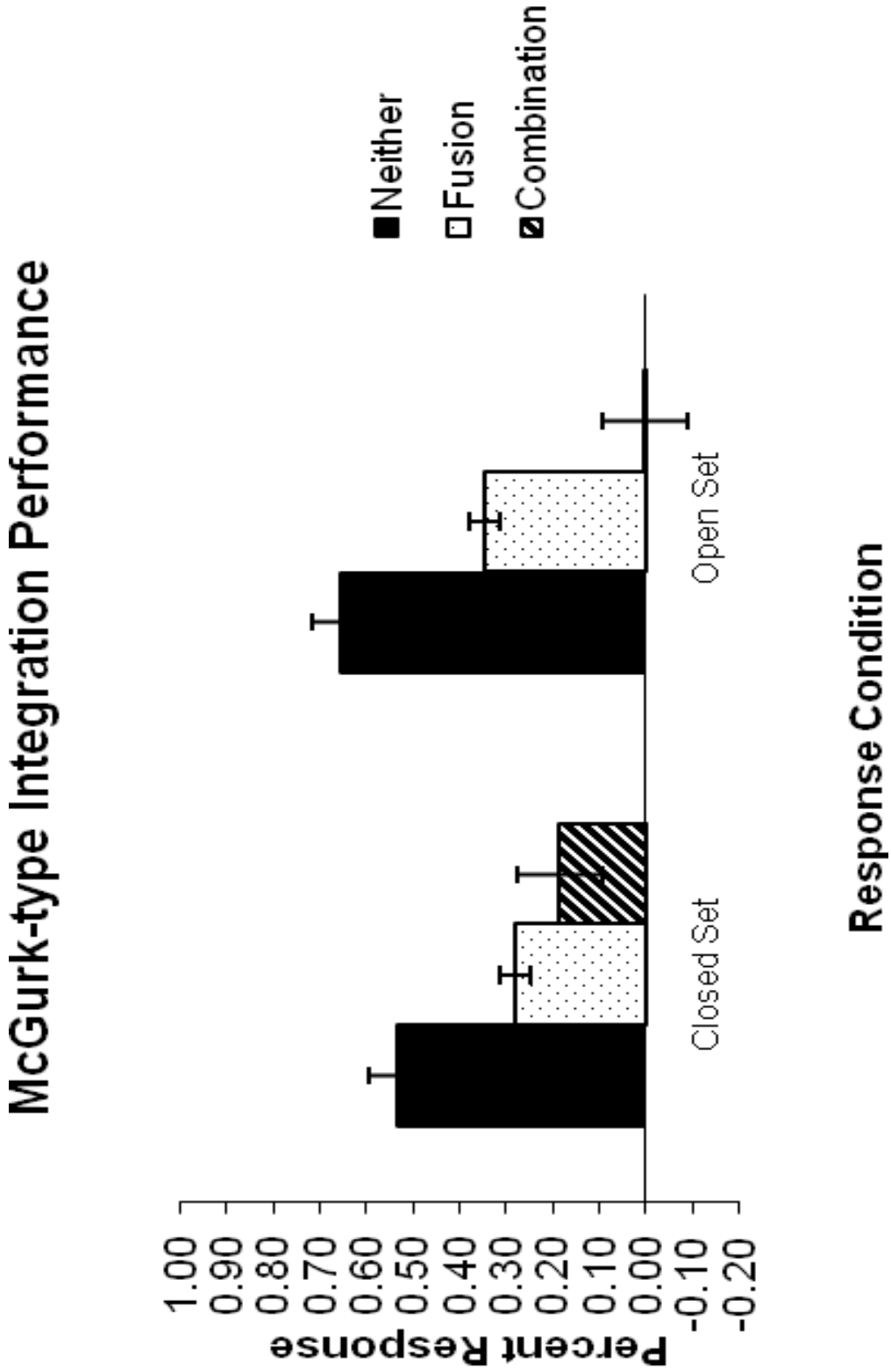
McGurk-type Integration Performance
Neither
Fusion
Combination
Percent Response
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
-0.10
-0.20
Closed Set
Open Set
Response Condition

Figure 5

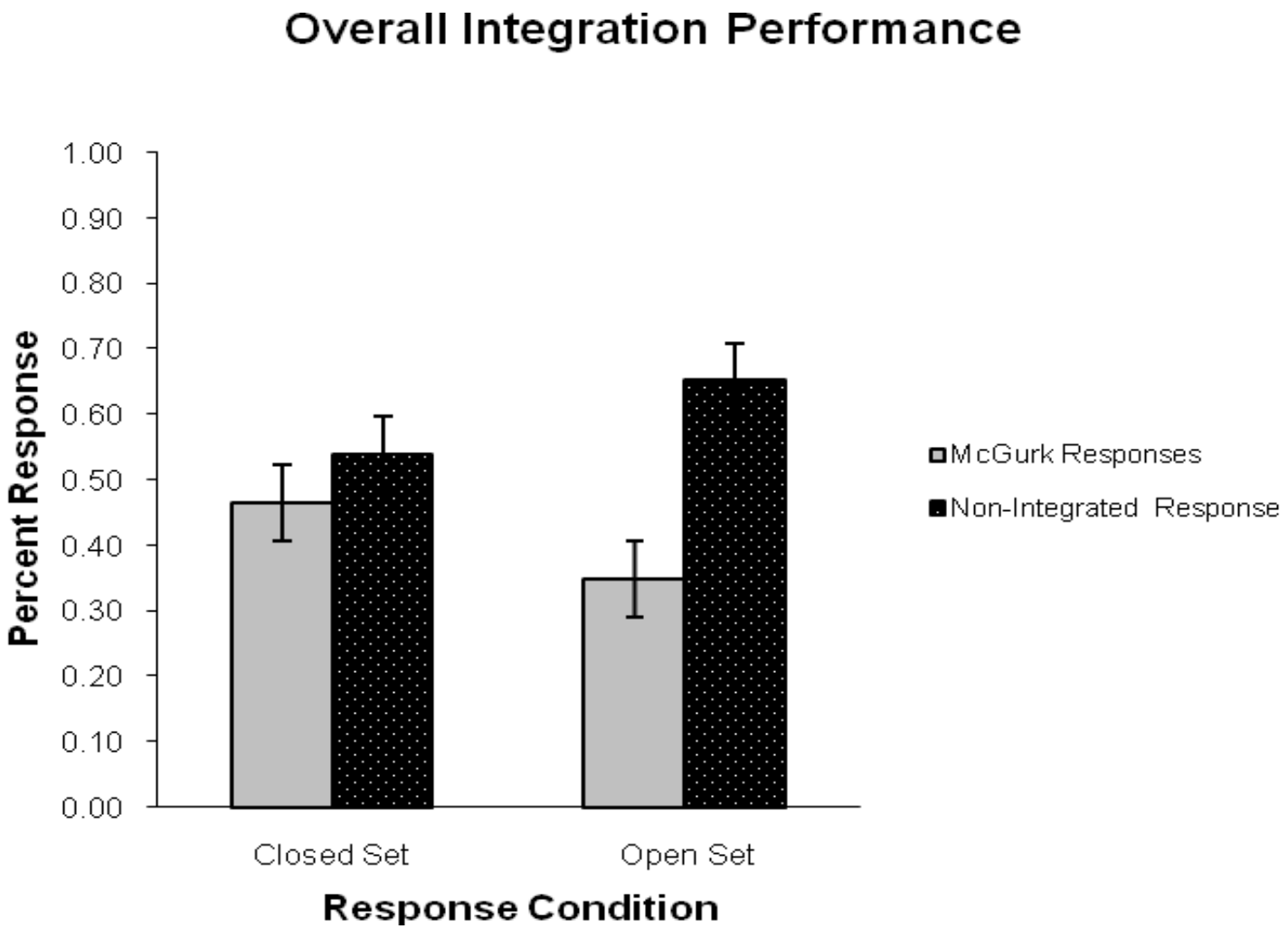
Overall Integration Performance
Percent Response
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
Closed Set
Open Set
Response Condition
McGurk Responses
Non-Integrated Response

Figure 6